

**KLASIFIKASI HALAMAN *WEB* MENGGUNAKAN METODE TF-IDF SVM
PADA SISTEM TERDISTRIBUSI *MULTINODE CLUSTER***

oleh

Kwang Dharma Saelau

NIM : 612012016



Skripsi

Untuk melengkapi salah satu syarat untuk memperoleh

Gelar Sarjana Teknik

Program Studi Teknik Elektro

Fakultas Teknik Elektronika dan Komputer

Universitas Kristen Satya Wacana

Salatiga

Agustus 2016



PERPUSTAKAAN UNIVERSITAS
UNIVERSITAS KRISTEN SATYA WACANA
Jl. Diponegoro 52 – 60 Salatiga 50711
Jawa Tengah, Indonesia
Telp. 0298 – 321212, Fax. 0298 321433
Email: library@adm.uksw.edu ; http://library.uksw.edu

PERNYATAAN TIDAK PLAGIAT

Saya yang bertanda tangan di bawah ini:

Nama : KWANG DHARMA SAELAU
NIM : 612012016 Email : 612012016@student.uksw.edu
Fakultas : TEKNIK ELEKTRONIKA DAN KOMPUTER Program Studi : TEKNIK ELEKTRO
Judul tugas akhir : KLASIFIKASI HALAMAN WEB MENGGUNAKAN METODE TF-IDF SVM
PADA SISTEM TERDISTRIBUSI MULTIMODE CLUSTER
Pembimbing : 1. BANU W. YOHANES, M. Comp Sc.
2. HARTANTO K.W, M.T.

Dengan ini menyatakan bahwa:

1. Hasil karya yang saya serahkan ini adalah asli dan belum pernah diajukan untuk mendapatkan gelar kesarjanaan baik di Universitas Kristen Satya Wacana maupun di institusi pendidikan lainnya.
2. Hasil karya saya ini bukan saduran/terjemahan melainkan merupakan gagasan, rumusan, dan hasil pelaksanaan penelitian/implementasi saya sendiri, tanpa bantuan pihak lain, kecuali arahan pembimbing akademik dan narasumber penelitian.
3. Hasil karya saya ini merupakan hasil revisi terakhir setelah diujikan yang telah diketahui dan disetujui oleh pembimbing.
4. Dalam karya saya ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali yang digunakan sebagai acuan dalam naskah dengan menyebutkan nama pengarang dan dicantumkan dalam daftar pustaka.

Pernyataan ini saya buat dengan sesungguhnya. Apabila di kemudian hari terbukti ada penyimpangan dan ketidakbenaran dalam pernyataan ini maka saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh karena karya saya ini, serta sanksi lain yang sesuai dengan ketentuan yang berlaku di Universitas Kristen Satya Wacana.

Salatiga, 4 JANUARI 2017


Tar KWANG DHARMA SAELAU



PERNYATAAN PERSETUJUAN AKSES

Saya yang bertanda tangan di bawah ini:

Nama : KWANG DHARMA SAELOU
NIM : 612012016 Email : 612012016@student.uksw.edu
Fakultas : FTEK Program Studi : TEKNIK ELEKTRO
Judul tugas akhir : KLASIFIKASI HALAMAN WEB MENGGUNAKAN METODE TF-IDF SVM
PADA SISTEM TERDISTRIBUSI MULTINODE CLUSTER

Dengan ini saya menyerahkan hak *non-eksklusif** kepada Perpustakaan Universitas – Universitas Kristen Satya Wacana untuk menyimpan, mengatur akses serta melakukan pengelolaan terhadap karya saya ini dengan mengacu pada ketentuan akses tugas akhir elektronik sebagai berikut (beri tanda pada kotak yang sesuai):

- ☒ a. Saya mengizinkan karya tersebut diunggah ke dalam aplikasi Repositori Perpustakaan Universitas, dan/atau portal GARUDA
- ☐ b. Saya tidak mengizinkan karya tersebut diunggah ke dalam aplikasi Repositori Perpustakaan Universitas, dan/atau portal GARUDA**

* Hak yang tidak terbatas hanya bagi satu pihak saja. Pengajar, peneliti, dan mahasiswa yang menyerahkan hak non-eksklusif kepada Repositori Perpustakaan Universitas saat mengumpulkan hasil karya mereka masih memiliki hak copyright atas karya tersebut.

** Hanya akan menampilkan halaman judul dan abstrak. Pilihan ini harus dilampiri dengan penjelasan/ alasan tertulis dari pembimbing I dan diketahui oleh pimpinan fakultas (dekan/kaprodi).

Demikian pernyataan ini saya buat dengan sebenarnya.

Salatiga, 5 JANUARI 2017

1956

Mengetahui,

BANU W. YOHANES, M.Com.Sc.

Tanda tangan & nama terang pembimbing I

KWANG DHARMA SAELOU

Tanda tangan & nama terang mahasiswa

HARTANTO KW, M.T.

Tanda tangan & nama terang pembimbing II

PERNYATAAN BEBAS PLAGIAT

Saya, yang bertanda tangan di bawah ini:

NAMA : Kwang Dharma Saelau

NIM : 612012016

JUDUL : KLASIFIKASI HALAMAN *WEB* MENGGUNAKAN
METODE TF-IDF SVM PADA SISTEM TERDISTRIBUSI
MULTINODE CLUSTER

Menyatakan bahwa skripsi tersebut di atas bebas plagiat. Apabila ternyata ditemukan unsur plagiat di dalam skripsi saya, maka saya bersedia mendapatkan sanksi apa pun sesuai dengan aturan yang berlaku.

Salatiga, September 2016



Kwang Dharma Saelau

1956

**KLASIFIKASI HALAMAN *WEB* MENGGUNAKAN METODE TF-IDF SVM
PADA SISTEM TERDISTRIBUSI *MULTINODE CLUSTER***

oleh
Kwang Dharma Saelau
NIM : 612012016


Skripsi ini telah diterima dan disahkan
Untuk melengkapi salah satu syarat memperoleh
Gelar Sarjana Teknik
dalam
Konsentrasi Teknik Telekomunikasi
Program Studi Teknik Elektro
Fakultas Teknik Elektronika dan Komputer
Universitas Kristen Satya Wacana
Salatiga

Disahkan oleh

Pembimbing I


Banu W. Yohanes, M.CompSc.
Tgl. 15-9-2016

Pembimbing II


Hartanto K.W, M.T.
Tgl. 14-8-2016

PERNYATAAN BEBAS PLAGIAT

Saya, yang bertanda tangan di bawah ini:

NAMA : Kwang Dharma Saelau
NIM : 612012016
JUDUL : KLASIFIKASI HALAMAN *WEB* MENGGUNAKAN
METODE TF-IDF SVM PADA SISTEM TERDISTRIBUSI
MULTINODE CLUSTER

Menyatakan bahwa skripsi tersebut di atas bebas plagiat. Apabila ternyata ditemukan unsur plagiat di dalam skripsi saya, maka saya bersedia mendapatkan sanksi apa pun sesuai dengan aturan yang berlaku.

Salatiga, September 2016



Kwang Dharma Saelau

INTISARI

Dengan semakin bertambah besarnya jumlah data dan halaman *web* di internet, semakin diperlukan pula sebuah *web crawler* yang memiliki efisiensi dan ketepatan dalam mencari sumber informasi yang diharapkan. Salah satunya adalah dengan memakai *focused web crawler*.

Dalam skripsi ini diusulkan sebuah *focused web crawler* yang diimplementasikan pada sebuah sistem terdistribusi Hadoop *Multinode Cluster* dengan metode TF-IDF SVM. Perancangan *focused web crawler* dengan sistem ini untuk mempercepat penelusuran halaman *web*, meningkatkan presisi dan akurasi topik serta meringankan proses kerja penelusuran.

Pada pengujian topik halaman *web* yang ditelusuri, digunakan 10 topik yaitu *smartphone*, banjir, bisnis, NBA(basket), olimpiade, pendidikan, pokemon, Portugal, sepak bola dan puasa. *Crawler* yang dirancang ini dapat menelusuri halaman *web* dengan kecepatan rata-rata 8,698 detik per halaman. Kemudian *crawler* ini juga memiliki *precision* 86,37% dan *recall* 66,68%.

Mengetahui,

Mengesahkan,

Penyusun,

1956

Dr. Iwan Setyawan
Dekan

Banu W. Yohanes, M.CompSc.
Pembimbing I

Kwang Dharma S.

ABSTRACT

With increasingly large amounts of data and web pages on the internet, the more necessary is also a web crawler that have high efficiency and accuracy in finding resources is expected. One of them is to use a focused web crawler.

In this thesis proposed a dedicated web crawler which is implemented in a distributed system Multinode Hadoop Cluster with TF-IDF SVM method. Design a dedicated web crawler with this system to speed up searches web pages, improving the precision and accuracy of topics as well as ease the work process search.

On testing the topic of the web pages searched, used 10 areas including smartphones, flooding, businesses, NBA (basketball), Olympic games, education, pokemon, Portugal, football and fasting. Crawler designed it can browse web pages with an average speed of 8,698 seconds per page. Then the crawler also has precision 86,37% and 66,68% recall.

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas penyertaan-Nya selama ini, sehingga penulis dapat menyelesaikan perancangan serta penulisan tugas akhir skripsi ini sebagai syarat kelulusan di Fakultas Teknik Elektronika dan Komputer Universitas Kristen Satya Wacana.

Pada kesempatan ini penulis juga berterima kasih kepada berbagai pihak yang baik secara langsung maupun tidak langsung telah membantu penulis dalam menyelesaikan skripsi ini :

1. Tuhan Yesus yang selalu menyertai, membimbing, serta memberikan jalan terbaik sehingga penulis dapat menyelesaikan skripsi ini.
2. Kedua orang tua tercinta yang sudah membesarkan dan mendidik penulis dengan kasih sayang serta ajaran-ajaran baik yang sangat berarti.
3. Bapak Banu W. Yohanes, M.CompSc. dan Bapak Hartanto K.W, M.T. sebagai pembimbing I dan pembimbing II, terima kasih atas bimbingan dan saran yang telah diberikan kepada penulis selama mengerjakan skripsi ini.
4. Kakak penulis Kak Liang Arta, S.T. yang selalu mendukung penulis agar cepat menyelesaikan tugas akhir.
5. Staff pengajar dan Tata Usaha yang banyak memberikan pelajaran berharga.
6. Keluarga besar 2012 sebagai teman seperjuangan yang selalu memberi dukungan kepada penulis.
7. Teman-teman terdekat Adi, Samuel, Fandy, Keenan, Yohanes, Raynaldy, Vallicano, dan Feliks yang telah bersama-sama menghadapi masa perkuliahan.
8. Berbagai pihak yang tidak dapat dituliskan satu persatu, penulis ucapkan terima kasih.

Penulis menyadari skripsi ini masih banyak kekurangan dan jauh dari kesempurnaan. Maka dari itu melalui kata pengantar ini penulis sangat terbuka menerima kritik serta saran yang membangun sehingga penulis dapat memperbaiki kekurangan yang ada. Akhir kata penulis sangat berharap agar skripsi ini dapat memberikan manfaat dan kontribusi terhadap perkembangan teknologi di masa yang akan datang.

Salatiga, 25 Agustus 2016

Penulis

DAFTAR ISI

INTISARI	i
ABSTRACT.....	ii
KATA PENGANTAR	iii
DAFTAR ISI.....	iv
DAFTAR GAMBAR.....	vi
DAFTAR TABEL.....	vii
BAB I PENDAHULUAN.....	1
1.1. Tujuan.....	1
1.2. Latar Belakang.....	1
1.2.1. Pendahuluan.....	1
1.2.2. Permasalahan	1
1.3. Spesifikasi Sistem	3
1.4. Sistematika Penulisan.....	3
BAB II DASAR TEORI	4
2.1. <i>Web Crawler</i>	4
2.1.1. Definisi.....	4
2.1.2. <i>Focused Web Crawler</i>	5
2.2. <i>Support Vector Machine (SVM)</i>	5
2.3. <i>Precision dan Recall</i>	7
2.4. Sistem Terdistribusi	7
2.5. <i>Cluster</i>	9
2.5.1. Definisi.....	9
2.5.2. <i>Master / Slave</i>	9
2.6. Apache Hadoop	9
2.6.1. Definisi.....	9
2.6.2. <i>Hadoop Distributed File system (HDFS)</i>	10
2.6.3. Apache Hadoop YARN	10

BAB III PERANCANGAN	12
3.1. Sistem Terdistribusi Hadoop	12
3.2. Persiapan <i>Environment</i> untuk Hadoop	13
3.3. Melakukan Pengaturan Hadoop untuk <i>Multinode Cluster</i>	15
3.4. Perancangan Database	17
3.5. <i>Focused Web Crawler</i> dengan TF-IDF SVM.....	18
BAB IV PENGUJIAN DAN ANALISISNYA	22
4.1. Cara Pengujian.....	22
4.2. Pengujian dan analisa <i>Crawler</i> pada Sistem Terdistribusi Hadoop.....	23
4.3. Perbandingan <i>Crawler</i> SVM TF-IDF dengan GA <i>Crawler</i>	28
BAB V KESIMPULAN DAN SARAN	29
5.1. Kesimpulan	29
5.2. Saran	29
DAFTAR PUSTAKA	30
LAMPIRAN A WEB INTERFACE HADOOP	31
LAMPIRAN B TABEL-TABEL HASIL PENGUJIAN	32

DAFTAR GAMBAR

Gambar 2.1. Arsitektur sebuah <i>Focused Web Crawler</i>	5
Gambar 2.2. Contoh hasil pencarian.....	7
Gambar 2.3. Aplikasi YARN.....	10
Gambar 2.4. YARN menjalankan sebuah aplikasi	11
Gambar 3.1. Blok Diagram Arsitektur <i>Web Crawling</i> memakai Sistem Terdistribusi	12
Gambar 3.2. Pemberitahuan mengenai Hadoop yang terinstall.....	14
Gambar 3.3. ERD Database <i>Crawler</i>	18
Gambar 3.4. Tabel-Tabel Database <i>Crawler</i> dengan MySQL	18
Gambar 3.5. Tahapan kerja <i>Focused Web Crawler</i> dengan SVM TF-IDF	18
Gambar 3.6. Flowchart Ekstraksi fitur menjadi model TF-IDF	19
Gambar 3.7. Flowchart Ekstraksi fitur menjadi model TF-IDF (lanjutan).....	20
Gambar A.1. Tampilan status node manager pada <i>web interface</i>	31
Gambar A.2. Tampilan status datanode pada <i>web interface</i>	31

DAFTAR TABEL

Tabel 1.1. Tabel perbandingan sistem yang diajukan penulis dengan sistem serupa.....	2
Tabel 4.1. Rangkuman Hasil Pengujian Topik Pertama.....	23
Tabel 4.2. Rangkuman Hasil Pengujian Topik Kedua.....	24
Tabel 4.3. Rangkuman Hasil Pengujian Topik Ketiga	24
Tabel 4.4. Rangkuman Hasil Pengujian Topik Keempat.....	25
Tabel 4.5. Rangkuman Hasil Pengujian Topik Kelima	25
Tabel 4.6. Rangkuman Hasil Pengujian Topik Keenam.....	25
Tabel 4.7. Rangkuman Hasil Pengujian Topik Ketujuh	26
Tabel 4.8. Rangkuman Hasil Pengujian Topik Kedelapan	26
Tabel 4.9. Rangkuman Hasil Pengujian Topik Kesembilan	27
Tabel 4.10. Rangkuman Hasil Pengujian Topik Kesepuluh.....	27
Tabel 4.11. Perbandingan <i>Crawler</i> SVM TF-IDF dengan GA <i>Crawler</i>	28
Tabel B.1. Hasil penelusuran Topik <i>Smartphone</i> pada <i>Database</i> tabel Frontiers	32
Tabel B.2. Hasil penelusuran Topik <i>Smartphone</i> pada <i>Database</i> tabel Pages	39